ČASOPIS PRO MODERNÍ FILOLOGII 99, 2017, Č. 2, S. 245–260

Databáze překladových ekvivalentů Treq[1]

Michal Škrabal (Praha) – Martin Vavřín (Praha)

THE TRANSLATION EQUIVALENTS DATABASE (TREQ)
The aim of the paper is to introduce a tool that has recently been developed at the Institute of the Czech National Corpus, the Treq Translation Equivalents Database, and to explore its possible uses. These range from simple, one-shot probes while searching for an equivalent expression for a target language to more sophisticated and elaborate corpus-assisted translations. A significant advantage of Treq is the possibility of clicking on any equivalent and immediately verifying its individual occurrences in context — and thus being able to more easily distinguish relevant translation candidates from misleading ones. This utility, which is based on data stored in the InterCorp parallel corpus, is continually being upgraded and enriched with new functions (the recent integration of multi-word units, adding English as the primary language of the dictionaries, an improved interface, etc.), and the accuracy of results is growing as the volume of data continually increases.

KEYWORDS
InterCorp, Treq, translation equivalents, alignment

KLÍČOVÁ SLOVA
InterCorp, Treq, překladové ekvivalenty, zarovnání

1. KORPUS INTERCORP

Korpus InterCorp (IC) je rozsáhlý paralelní synchronní korpus budovaný v Ústavu Českého národního korpusu již od roku 2005. V posledních letech dochází k systematickému rozšiřování korpusu každý rok a od roku 2013 (verze 6) jsou předchozí verze korpusu dostupné v rozhraní KonText. IC je složen z několika částí: ručně korigovaných beletristických textů s kontrolovaným zarovnáním (tzv. *jádro*) a tzv. *kolekcí*, jež obsahují texty zpracovávané nikoliv manuálně, ale automaticky. Jde o tyto typy textů:

1 Tento článek vznikl při realizaci projektu Český národní korpus (LM2015044) financovaného Ministerstvem školství, mládeže a tělovýchovy v rámci aktivity Projekty velkých infrastruktur pro VaVaI.
Za cenné připomínky k článku chceme poděkovat svým kolegům Michalu Křenovi a Alexandru Rosenovi, jenž se zároveň podílel na návrhu a realizaci celého projektu Treq. Dále za podněty k vývoji tohoto nástroje vděčíme Elżbietě Kaczmarské, za technickou pomoc děkujeme Ondřeji Bojarovi a Davidu Marečkovi a za výpomoc s grafickým zpracováním kolegovi Janu Kockovi. V neposlední řadě patří naše poděkování též dvojici anonymních recenzentů tohoto článku.

— právní texty Evropské unie z korpusu Acquis Communautaire,
— publicistické články a zpravodajství z webových stránek Project Syndicate a VoxEurop (dříve PressEurop),
— záznamy jednání z Evropského parlamentu z let 2007–2011 z korpusu Europarl,
— filmové titulky ze serveru Open Subtitles.

Aktuální IC v9 obsahuje kromě tzv. pivotního jazyka[2] — češtiny — dalších 39 jazyků, které jsou ovšem co do velikosti i složení zastoupeny nerovnoměrně: největším částí korpusu je anglická složka (necelých 120 milionů slovních tvarů, z toho přes 21 milionů slov tvoří jádro), největším jádrem disponuje němčina (přes 31 milionů slov), ale některé jazyky ručně zpracované jádro vůbec nemají a obsahují pouze balíčky (např. vietnamština o celkové velikosti necelých 1,5 milionu slov, tvořených výhradně filmovými titulky, aj.). Texty u více než poloviny jazyků jsou opatřeny morfologickou anotací (23 z 39) a lemmatizovány (20 z 39). Celková velikost IC v9 činí více než 1,2 miliardy slovních tvarů, resp. přes půldruhé miliardy tokenů.[3]

2. DATABÁZE TREQ A MOŽNOSTI JEJÍHO VYUŽITÍ

Nástroj Treq je poměrně nový, v odborném tisku dosud nepopsaný (s dílčí výjimkou Kaczmarska — Rosen, 2015); jeho prvotní verze (0.1 alpha) pochází ze září 2014,[4] rychle si však získává oblibu mezi uživateli,[5] zejména pro svou jednoduchost a přímočarost. Čerpá z dat IC, která se zpracovávají pomocí automatických nástrojů, popsaných v následující kapitole; výsledkem je databáze překladových ekvivalentů, která je uživatelům volně přístupná[6] na webové stránce https://treq.korpus.cz.

Kromě nejrozšířenějšího způsobu užití, kdy hledáme nějaký ekvivalent pro výraz ve výchozím jazyce (možná extenze tohoto základní použití je popsána dále, v oddílu 5), lze na protějšky nabízené Treqem pohlížet i jako na potenciální slovníkové ekvivalenty. Lexikografům se tak dostává do rukou nástroj, jenž jim v jediném okamžiku nabídne soupis kandidátů na protějšky v cílovém jazyce i s frekvencí těchto ekvivalencí (absolutní i vyjádřenou procentuálně); často přitom platí, že čím je tato frekvence vyšší, tím pravděpodobněji je daný kandidát funkčním ekvivalentem. Podstatnou výhodou je

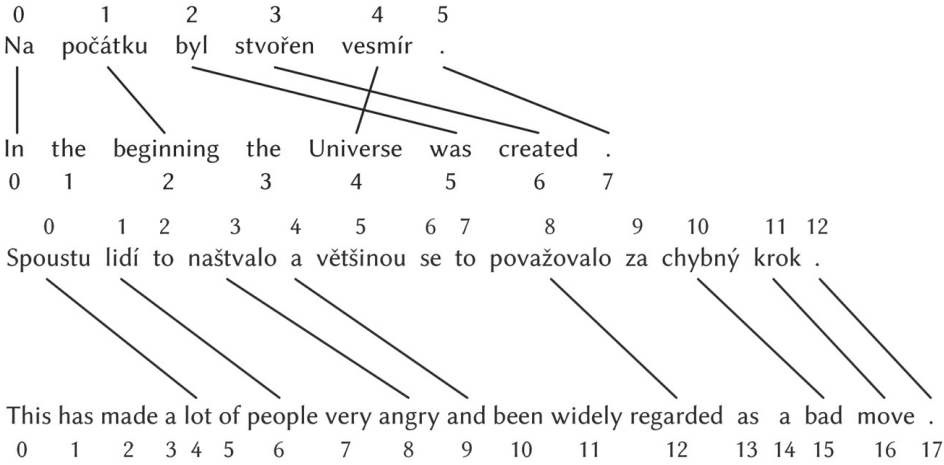
2 Pivotním jazykem se rozumí centrální jazyk v daném paralelním korpusu, vůči němuž jsou všechny texty v korpusu zarovnávány.
3 Přesné složení korpusu najdete na stránkách: http://wiki.korpus.cz/doku.php/cnk:intercorp:verze9. Obecně o InterCorpu viz Čermák — Rosen, 2012, či Rosen, 2016.
4 K verzím podrobněji viz Info o verzi na stránce https://treq.korpus.cz/.
5 Za rok 2016 bylo na portálu www.korpus.cz evidováno přes 719 tisíc dotazů; nejhojněji využívaným nástrojem je KonText (s více než 85 %), následovaný právě databází Treq (více než 70 tisíc dotazů, tj. bezmála 200 denně, což představuje necelých 10 % z celkového počtu zadaných dotazů).
6 Samotné výsledky hledání jsou přístupné i bez registrace, stejně tak lze přejít do rozhraní KonText a vidět nabízené ekvivalenty v jejich přirozeném prostředí. Pro plnohodnotnou práci s těmito daty v KonTextu je však nezbytné být přihlášen ke svému uživatelskému účtu.

možnost ověřit si jednotlivé realizace kteréhokoliv z nich přímo v kontextu pomocí hypertextového odkazu — a snáze tak odlišit relevantní kandidáty od těch zavádějících.[7]

3. ZPRACOVÁNÍ DAT[8]

Při přípravě dat pro databázi Treq nejprve vybereme z celého korpusu dané jazykové verze IC pouze věty, které jsou k češtině zarovnány v poměru 1:1. Výhradně jednoduchá zarovnání používáme proto, že obvykle bývají spolehlivější; zvláště v případě automaticky zarovnaných textů tak můžeme předejít zanesení potenciálních chyb.

Následuje automatické zarovnání po slovech v rámci těchto vět pomocí programu GIZA++ (Och — Ney, 2003).[9] Dosavadní verze Trequ (0.1–1.1) využívaly zarovnání pomocí metody *intersection*; vznikají tak pouze dvojice, v nichž jedno slovo odpovídá jednomu ekvivalentu, např.:



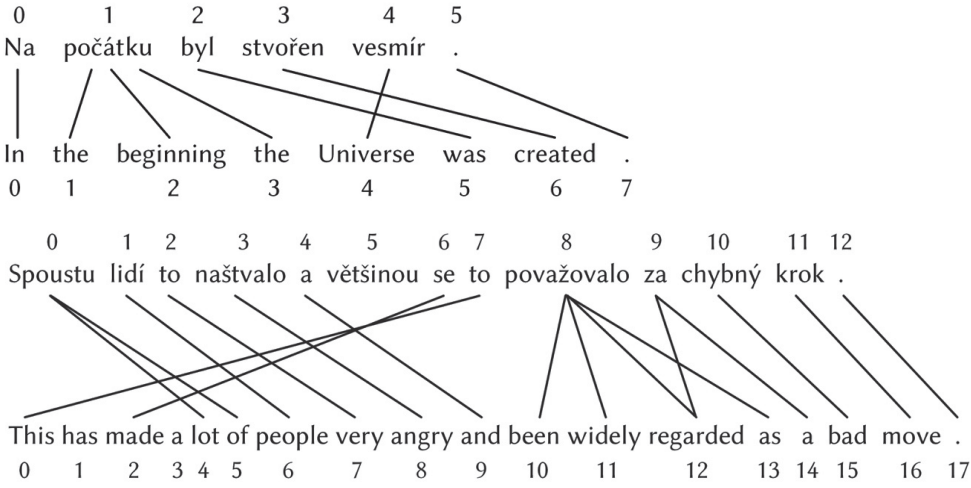
7 Vytěžování dat z paralelních korpusů pro lexikografické účely je logickým postupem, tkvícím v samotné povaze těchto dat. Přehled o možnostech extrakce dvoujazyčných slovníků z paralelních a srovnatelných korpusů podává Sharoff et al. (2013), srov. též diplomovou práci J. Tiedemanna (2000). Dílčí pokusy v tomto ohledu byly podniknuty také v českém prostředí, a to v případě angličtiny (Čmejrek — Cuřín, 2001; srov. též Čmejrek, 1998, a Cuřín, 1998), litevštiny (Skoumalová, 2008) či chorvatštiny (Jirásek, 2011). Kupř. poslední dva jmenovaní autoři se shodují na tom, že slovníky takto automaticky extrahované jsou pouhým východiskem pro následnou slovníkářskou práci, mohou ji však lexikografovi nemálo ulehčit. To ostatně potvrzuje i naše vlastní zkušenost: Treq je využíván při vytváření lotyšsko-českého slovníku (Škrabal, 2016b).
— Potenciál databáze Treq pro lexikografické účely by si bezesporu zasloužil samostatnou studii, v tomto článku jej zmiňujeme pouze okrajově.

8 Srov. proces vzniku „statistických překladových slovníků" (Kovář — Baisa — Jakubíček, 2016, s. 343n.).

9 Konkrétní instalace viz https://github.com/moses-smt/mgiza/tree/master/mgizapp. Využit byl též pomocný skript od Ondřeje Bojara.

Tzn. že v první větě první slovo ve zdrojovém jazyku (0) odpovídá prvnímu slovu v jazyku cílovém (0), druhé slovo (1) odpovídá třetímu slovu (2) atd. (srov. Rosen — Adamová — Vavřín, 2014; Kaczmarska — Rosen, 2015, s. 164–165).

Pro novou verzi (2.0) jsme se rozhodli kromě těchto jednoduchých zarovnání navíc využít metodu *grow-diag-final-and*, která umožňuje vytvářet i komplikovanější zarovnání více slov na obou stranách překladu.[10] Takové zarovnání pak může vypadat třeba takto:



Oproti případu výše tu druhé slovo ve zdrojovém jazyku (1) neodpovídá pouze třetímu (2), ale též druhému a čtvrtému (1, 3) slovu v jazyku cílovém atd.

Z takovéhoto zarovnání následně vybíráme pomocí jednoduchého skriptu co největší množství kombinací slov, které toto zarovnání umožňuje (viz též příklad extrahovaných ekvivalentů níže).

V obou případech jsou zarovnané dvojice slov setříděny a sečteny, výsledky automatické extrakce však už nejsou nijak revidovány a uživateli jsou poskytnuty formou seznamu nalezených ekvivalentů zadaného výrazu, doplněných o absolutní a relativní frekvenci.

V jakém poměru jsou frekvence nalezené v KonTextu s těmi zobrazovanými Treqem, ukazuje Tabulka 1. Ta vyčísluje různé typy dat v jednotlivých fázích jejich zpracování pro Treq z česko-anglické složky IC v9 (víceslovná varianta).

10 Jednotlivé metody zarovnávání slov popisují a srovnávají např. Mareček, 2009, či Girgzdis et al., 2014.

Fáze zpracování	Výsledná data		Počet (v tisících)						
			Core	Subtitles	Acquis	Europarl	VoxEurop	Syndicate	Celkem
0. Vstup	Pozice v angličtině		25 149	66 790	29 626	17 384	3 123	4 387	146 458
	Věty v angličtině		1 510	9 211	1 426	681	152	190	13 171
1. Zarovnání vět 1:1	Zarovnané věty	lemmata	1 267	6 955	1 251	656	127	180	10 437
		tvary	1 267	6 955	1 254	656	127	180	10 440
2. Zarovnání slov	Nalezené ekvivalenty	lemmata	15 785	41 189	19 344	12 812	1 670	3 352	94 153
		tvary	15 538	41 445	19 656	12 899	1 598	3 344	94 479
3. Vytvoření slovníku	Položky slovníku	lemmata	3 235	6 697	1 441	1 213	547	550	13 682
		tvary	4 639	9 276	2 056	1 946	670	873	19 460
4. Vyčištění slovníku	Položky slovníku	lemmata	2 775	5 375	1 133	1 061	461	458	11 263
		tvary	3 966	7 146	1 722	1 760	566	750	15 909

TABULKA 1. Zpracování dat pro česko-anglický slovník

Po dílčích krocích lze sledovat postupný úbytek dat, která jsou ve výsledném slovníku použita. V prvním kroku použijeme pouze zarovnání vět 1:1 — tím přijdeme o 20,7 % vět.[11] Následně se vyberou na základě zarovnání z programu Giza++ jedno- a víceslovné ekvivalenty. Vztah mezi velikostí původního korpusu a počtem vyextrahovaných ekvivalentů však nelze jasně předvídat, zvláště pak u víceslovných ekvivalentů, kde vznikají nejrůznější kombinace stejných slov (viz tučně vysázené dvojice níže). Takto by např. vypadal abecedně řazený soupis česko-anglických párů extrahovaných z druhé příkladové věty:

a – and
chybný — bad
krok — move
lidí — people
naštvalo — angry
považovalo — been widely regarded as
považovalo za — been widely regarded as
považovalo za — regarded a
se — made
Spoustu — lot of
to — This
to — very
za — regarded a
. — .

11 V budoucnu lze experimentovat rovněž s neparitním zarovnáváním. Další možné plány jsou naznačeny v samotném závěru této studie.

Ve třetím kroku se v rámci celého textu sečtou řádky, které jsou stejné na obou stranách zarovnání. Tak získáme seznam a frekvenci ekvivalentů. Nakonec, v závěrečném kroku, vyřadíme všechny protějšky obsahující interpunkci, čímž obdržíme finální verzi slovníku. U všech jazykových párů, kde je k dispozici lemmatizace na obou stranách zarovnání, aplikujeme stejný postup i na lemmatizovanou podobu dat (*na počátek být stvořit vesmír . — in the beginning the universe be create .*).

4. ROZHRANÍ

Přístup k takto extrahovaným datům pak zprostředkovává rozhraní Treq. Ve výchozím nastavení jsou protějšky hledaného výrazu řazeny sestupně podle četnosti; jejich relativní frekvence je uživatelovým primárním vodítkem: čím častěji se ekvivalent hledaného výrazu vyskytl ve srovnání s ostatními ekvivalenty, tím vyšší je pravděpodobnost, že je funkční. U dostatečně velkých a žánrově pestrých korpusů má zajisté smysl uvádět frekvenci ekvivalentních párů zvlášť pro různé typy textů (beletrie, publicistika, právnické texty, filmové titulky suplující mluvený jazyk), tuto možnost Treq rovněž nabízí (pole *Omezit na*).

Počínaje verzí 2 je možné do dotazovacího okénka zadávat víceslovné jednotky — a to v obou směrech — a očekávat výsledky jak jednoslovné, tak víceslovné, přičemž uživatelé mají na výběr mezi oběma možnostmi (spolu s jinými eventualitami, např. nerozlišováním mezi velkými a malými písmeny). Možnosti Trequ se tím podstatně rozšiřují: např. pro kombinaci angličtina-čeština lze nyní vyhledávat mnohé třídy slov: frázová slovesa, diskursní částice (*discourse markers*), fráze v obecném smyslu aj., v opačném směru např. reflexivní slovesa. Mimoto nynější výsledky více odpovídají reálnému jazykovému stavu: ekvivalenci lexémů ve zdrojovém a cílovém jazyce nelze pochopitelně omezovat na „ideální" poměr 1:1.

S víceslovnými výrazy získala na naléhavosti potřeba zakomponovat také dotazovací jazyk, který by umožňoval dotazy pomocí zástupných výrazů;[12] dosud Treq vyhledával pouze přesné řetězce znaků. Kromě toho byl pro verzi 2 rozšířen primární jazyk slovníků z dosavadní češtiny na angličtinu: vedle oboustranných česko-cizojazyčných slovníků byly z dat v IC automaticky vygenerovány oboustranné slovníky anglicko-cizojazyčné. Možnost využívat Treq se tak otevírá mnohem širšímu publiku než dosud, uživatelé už nejsou limitováni nutností ovládat češtinu. Teoreticky lze v budoucnu rozšířit primární jazyk na kterýkoliv jazyk v IC zastoupený; v tomto ohledu je nutno řídit se především zájmem a potřebami uživatelů.

12 Treq je postaven na databázovém systému MySQL, který využívá knihovnu regulárních výrazů vytvořenou Henrym Spencerem a vyhovující standardu POSIX a (viz např. https://garyhouston.github.io/regex/). Detailnější popis dotazovacího jazyka najdete v manuálu k nástroji Treq: http://wiki.korpus.cz/doku.php/manualy:treq.

5. DATABÁZE TREQ V PRAXI

Další možné využití, jež mělo zároveň posloužit jako jakási „zatěžkávací zkouška" nové verze Trequ (v té době dosud jen interně dostupného prototypu), bylo prezentováno v rámci kolokvia KOLT 2016, konaného loni v listopadu (Škrabal, 2016a). Mottem daného kolokvia byl citát S. Johanssona (2007, s. 1):

(0) *It has often been said that, through corpora, we can observe patterns in language which we were unaware of before* […]. *My claim is that this applies particularly to multilingual corpora. We can see how languages differ, what they share and — perhaps eventually — what characterises language in general.*

Protože neexistuje překlad Johanssonovy monografie do češtiny, požádali jsme své čtyři kolegy, vesměs anglisty nebo překladatele z/do angličtiny, o *ad hoc* převod těchto tří vět.

(1) *Často se říká, že skrz korpusy je možné v jazyce nahlédnout pravidelnosti, které nám doposud zůstávaly skryty. Podle mě toto tvrzení platí dvojnásob v případě korpusů vícejazyčných. Vyčteme z nich, v čem se jazyky liší, co sdílejí, a někdy třeba i obecné vlastnosti jazyka jako takového.*

(2) *Již bylo mnohokrát řečeno, že prostřednictvím korpusů můžeme v jazyce objevit vzorce, kterých jsme si do té doby nebyli vědomi. Domnívám se, že toto tvrzení platí dvojnásob o vícejazyčných korpusech. Můžeme v nich pozorovat, jak se jazyky liší, co sdílejí a možná nakonec i odhalit to, co charakterizuje jazyk jako takový.*

(3) *Často se říká, že prostřednictvím korpusů můžeme v jazyce sledovat trendy, kterých jsme si dříve nebyli vědomi. Já tvrdím, že to platí zejména pro vícejazyčné korpusy. Vidíme, jak se jazyky od sebe liší, co mají společného, a nakonec snad také to, co charakterizuje jazyk obecně.*

(4) *Často se tvrdí, že pomocí korpusu lze vysledovat jazykové vzorce, kterých si jinak nejsme vědomi. Já tvrdím, že toto platí zejména na korpusy vícejazykové. Vidíme na nich, jak se jazyky vzájemně liší, v čem jsou si podobné, a snad jejich pomocí nakonec zjistíme, co je pro jazyky charakteristické obecně.*

Variabilita je tu zjevná a pochopitelná, a s rostoucím počtem překladatelů by jistě nadále rostla, ale stejně tak bychom našli i opakující se překladatelská řešení; už na tomto vzorku jsou vedle odlišností patrné určité paralely, např. všichni čtyři zvolili sloveso *platit*, zato už každý s jinou předložkovou vazbou.

Podobně bychom mohli srovnat i překlady strojové, generované automatickými překladači. Vybrali jsme tři, budované na statistických základech: Google Translator (GT),[13]

13 https://translate.google.cz/.

Microsoft Translator (MT), integrovaný do textového editoru Word,[14] a překladač Moses, vyvíjený Ústavem formální a aplikované lingvistiky MFF UK.[15]

(5) *To často bylo říkal, že prostřednictvím korpusů, můžeme pozorovat vzory v jazyce, kterému jsme byli vědomi dříve. Můj požadavek je, že to platí zejména pro vícejazyčného korpusy. Můžeme vidět, jak jazyky se liší, co sdílejí a — možná nakonec — co charakterizuje jazykem obecně.*

(6) *Často bylo řečeno, že prostřednictvím korpusy, můžeme sledovat vzorce v jazyce, který jsme nevěděli před. Moje tvrzení je, že to platí zejména pro vícejazyčné korpusů. Můžeme vidět, jak jazyků liší, co sdílejí a — možná nakonec — co je charakteristické pro jazyk obecně.*

(7) *To často zaznělo, že prostřednictvím korpusů, můžeme pozorovat vzory v jazyce, kterému jsme byli vědomi. Žaloby je, že to platí zejména pro vícejazyčné corpora. Můžeme vidět, jak se liší jazyky, co sdílejí a — možná nakonec — co charakterizuje jazyk obecně.*

V zásadě lze tvrdit, že jde o obstojné výsledky, které pro rámcovou představu o smyslu daného textu víceméně postačí. Náš zájem však netkví v pasivním porozumění cizímu textu, ale v aktivním, tvůrčím procesu převodu textu z jazyka zdrojového do cílového; zajímá nás, nakolik v takovémto procesu mohou napomoci data z paralelního korpusu IC, ať už přes rozhraní KonText, či Treq. Na základě těchto dat jsme se pokusili vytvořit *ad hoc* překlad, který by se dal označit za *corpus-assisted*. Důležitá byla post-editační fáze překladu, kterou umožňují rovněž některé překladače (z námi uvedených GT): uživatel si na úrovni jednotlivých slov či kolokací vybírá alternativní překladatelská řešení, v případě anglicko-českého překladu jsou nabízeny kromě synonymních ekvivalentů i jiné pádové formy. Původní podobu si tak uživatel upravuje v souladu se svou představou o ideálním výsledku. Společným prvkem je tu jeho aktivní role: nezůstává jen pasivním příjemcem daného produktu, ale stává se činným post-editorem.

Daný citát jsme si pro své potřeby rozčlenili do menších ucelených jednotek.[16] Je nutno přiznat, že tato parcelace je do značné míry intuitivní a zdaleka ne jediná možná. Neskrýváme tu ani předběžnou znalost zdrojového jazyka, nejde nám o perspektivu badatelskou, ale spíše uživatelskou, o situaci, kdy se uživatel snaží tyto víceslovné jednotky, volněji či pevněji spojené, přeložit pomocí korpusových nástrojů.

14 Používáme verzi editoru Word 2007. Mimoto je MT k dispozici online na adrese https://www.bing.com/translator.

15 Dostupný online na http://lindat.mff.cuni.cz/services/moses/. Všechny tři strojové překlady jsou citovány k 3. 12. 2016.

16 Záměrně se tu vyhýbáme jakémukoliv bližšímu označení, byť se v anglickojazyčné odborné literatuře nabízejí mnohé, více či méně přesně definované, pojmy jako např. *lexical bundles*, *lexical chunks*, *lexical clusters*, *patterns*, *N-grams* aj. Srov. též přístup tzv. *Linear Unit Grammar*, který nabízejí J. Sinclair a A. Mauranenová (2006).

Nezbytná je přitom jeho obeznámenost (byť jen elementární) s daným zdrojovým jazykem, tak aby si z nabízených kandidátů na ekvivalenty překládaného výrazu mohl správně vybrat. V tomto ohledu je databáze Treq nástrojem sloužícím často k aktivaci pasivních znalostí či k potvrzení uživatelových domněnek — a jako takovou je třeba ji brát *cum grano salis*. V žádném případě ji nelze zaměňovat za regulérní slovník, měla by sloužit spíše jako jeho doplněk, který kýžené překladatelské řešení nenabízí automaticky, může však uživatele alespoň navést správným směrem. Oproti klasickým slovníkům odráží ve větší míře specifičnost některých případů mezijazykové ekvivalence, kromě frazeologičnosti např. též odlišnou slovnědruhovou platnost výrazu a jeho cizojazyčného ekvivalentu či jejich neparitní zarovnání (tj. v poměru 1:n / n:1). Zároveň nabízí nesrovnatelně bohatší exemplifikaci, všechny příklady nadto pocházejí z autentických, dále neupravovaných zdrojů.

Námi rozparcelované věty vypadají takto:

(0/1) *It* ***has often been said*** *that,* | ***through*** *corpora,* | ***we can observe*** | ***patterns in*** *language* | *which* ***we were unaware of*** *before.*

(0/2) ***My claim is*** *that* | ***this applies*** *particularly* ***to*** | ***multilingual corpora****.*

(0/3) ***We can see*** | ***how*** *languages* ***differ****,* | ***what*** *they* ***share*** | *and* — ***perhaps eventually*** — | ***what characterises*** *language* | ***in general****.*

Jak vidno, jednotlivé úseky se od sebe dosti liší: kromě své délky též sémanticky, strukturně, mírou ustálenosti aj. Netučně jsou vysázeny jednak spojovací výrazy, jednak lexikální proměnné, které mohou variovat (ovšem ne libovolně a bez jakýchkoliv omezení); zvýrazněno je pomyslné jádro daného sousloví, jež budeme hledat v (celém) korpusu InterCorp v9 — English (Klégr et al., 2016): v rozhraní KonText nejčastěji pomocí typu dotazu *Fráze* (jednotlivé výrazy též pomocí *Slovního tvaru*; lze pochopitelně vždy využít také nejobecnější typ dotazu, tj. *CQL*), v rozhraní Treq prostým vepsáním dané fráze či slova (vzhledem k flektivnosti češtiny volíme zarovnání po slovních tvarech, tj. nevybíráme volbu *Lemmata*[17]; zaškrtnutu máme naopak volbu nerozlišovat velikost písmen), s možností využití výše zmíněných zástupných výrazů. Obecně platí, že se vždy snažíme najít co největší část tohoto sousloví; a lze předpokládat, že počet dokladů bude klesat úměrně s délkou daného dotazu — většinou půjde o malé počty výskytů: jednotky či pár desítek.[18]

Potvrzuje se to hned v úvodním případě: vstupní frázi, resp. její část *has often been said* najdeme v KonTextu pouze 9krát, přičemž všechny překlady (včetně toho posledního, nejvolnějšího) se zdají být relevantní: *často se říká* (2), *často zaznělo* (2), po

17 Dlužno podotknout, že zarovnání po lemmatech může přinést výsledky odlišné, nikoliv však zásadně. Snadná změna v nastavení dotazu umožňuje uživateli experimentovat s jednotlivými mody a rozhodnout se, který mu vyhovuje více.

18 Považujeme za důležité znovu zopakovat, že to nemusí být *a priori* na škodu: naším cílem není žádná lingvistická či translatologická analýza, snažíme se vcítit do řadového uživatele překládajícího anglický text.

1 výskytu *říká se často*, [komentátoři] *často tvrdí*, *mnohokrát zaznělo*, *bylo mnohokrát řečeno*, *opakovaně jsem slyšel*. V Trequ se daný obrat nevyskytuje, je zapotřebí jej dále omezit na *has been said*; dotaz pak přinese tyto protějšky:

> *zaznělo* (31), *řečeno* (6), *bylo řečeno* (3), *již* (2), 13 protějšků s frekvencí 1, z toho tyto víceslovné: *jsme se*, *už řečeno*, *se již*, *jsem již*, *již nejednou*, *zaznělo zde*, *již řekl*, *zaznělo názorů*

Pro účely svého experimentu jsme vybrali ekvivalent *často se říká*. Dále se budeme primárně řídit frekvenčním hlediskem; tam, kde bude na výběr z více možností, aniž by jedna z nich výrazně převažovala, zvolíme takové řešení, které by působilo co nejústrojněji v kontextu celé věty.

Následující sousloví *through corpora* je zajímavé hned z několika hledisek. Jednak jde o předložkovou vazbu, kde tvar *corpora* může být nahrazen jiným (i rozvitým) jmenným výrazem, a protože nejde o ustálený obrat, hledání v IC jsme nuceni provést pomocí dvou dotazů, na obě části zvlášť. Frekvence obou lexémů je přitom diametrálně odlišná: zatímco první najdeme 71 334krát, druhý pouze v 5 výskytech, přičemž jen jedinkrát v hledaném významu, což je dáno vedle nedostatečné velikosti též žánrovým složením IC.[19] Zde se neobejdeme buď bez znalosti daného odvětví, nebo bez konzultace s jinou překladovou příručkou. Přednosti korpusového přístupu, a zejména potenciál nástroje Treq, se naopak vyjeví u lexémů s dostatečně vysokou frekvencí, kdy výsledný konkordanční seznam přesahuje možnosti manuální analýzy. Po vepsání výrazu *through* v Trequ (všechny volby jsou odškrtnuté) okamžitě dostaneme četné kandidáty na ekvivalenty:

> *prostřednictvím* (7139), *přes* (2583), *skrz* (1225), *pomocí* (861), *skrze* (744), *projít* (551), *díky* (436), *prošel* (288), *do* (284), *základě* (199) …

Obdobný postup je třeba uplatnit i v případě sousloví *multilingual corpora* ve větě 0/2, jež je jediným víceslovným termínem v celém Johanssonově citátu. V KonTextu najdeme celkem 129 výskytů přívlastku *multilingual* (ani jednou však v námi hledaném spojení). Omezíme-li se pak v Trequ na nabízené adjektivní ekvivalenty (pro větší přehlednost s přepnutím na zarovnávání podle lemmat), dostaneme tyto:

> *mnohojazyčný* (47), *vícejazyčný* (29), *vícejazykový* (1), *mnohajazyčný* (1), *osvícený* (1), *vícerojazyčných* (1), *multilingvních* (1)

Ke správné volbě je mnohdy nezbytné znát specifika dané oborové terminologie, případně se o nich poučit porovnáním kontextů, v nichž se jednotliví kandidáti vyskytují. V tomto případě se oba nejčastější ekvivalenty v českém prostředí užívají.[20]

19 Jde o situace, kdy automatické překladače pro nedostatek dat daný výraz nepřeloží a ponechají ho v originále (jako se to stalo v příkladové větě 7).

20 Chlumská (2014, s. 224) zavádí jako ekvivalent k *multilingual corpora* výraz *vícejazyčné korpusy*, Čermák — Kocek (2010) titulují IC označením *mnohojazyčný korpus*.

Obrat *we can observe* najdeme v IC 17krát, nabídka překladatelských řešení je velmi pestrá, srov.

> *můžeme pozorovat* (5), *pozorujeme* (3), *vidíme* (2), *můžeme sledovat* (1), *můžeme vidět* (1), *máme možnost vidět* (1), *vidíme* (1), *máme k dispozici* (1), *pozorovatelná* [část vesmíru] (1), v jednom případě nepřeloženo

— a dokonce najdeme i 5 výskytů v Trequ, s českými protějšky *vidíme, vidíme toho* a *pozorujeme*.

Strukturně obdobné sousloví *we can see* (z věty 0/3) je mnohem frekventovanější, v IC je doloženo 843krát, 27krát s následným výrazem *how*, nejčastějšími českými ekvivalenty jsou *vidíme* (5), *víme* (3), *můžeme vidět* (2), *je vidět* (2). V Trequ na dotaz *we can see* dostaneme 272 výsledků, nejčastějším je *vidíme* (184, cca 68 %), následují *uvidíme* (26), *zjistíme* (5), *pozorujeme* (3), *víme* (3) a další výrazy s frekvencí 2 či 1.

V úseku *patterns in language* považujeme za klíčová první dvě slova, ta jako frázi hledáme v KonTextu. Ze 71 výskytů jako protějšek lexému *patterns* nejčastěji (11krát) figuruje výraz *vzorce* (*vzory* 4krát, *vzorky* 3krát), *způsoby* (8krát), *obrazce* (4krát, jednou *obrazy*) a *modely* (3krát), jinak se české ekvivalenty opakují minimálně či vůbec. Treq poskytuje pouze 3 protějšky po jediném výskytu: *s hroznovým*, *s*, *principy*. Srov. nejčastější (též víceslovné) ekvivalenty lemmatu *pattern* nabízené Treqem:

> *vzorec* (257), *struktura* (166), *vzor* (158), *model* (123), *schéma* (75), *obrazec* (67), *tok* (53), *vzorek* (51), *chování* (35), *typ* (32), *způsob* (29), *gilošovaný vzor* (18), *šablona* (13), *vibrační vzorek* (12), *skladba* (11), *postup* (11), *uspořádání* (10), *systém* (10)

Pro účely svého překladu volíme nejfrekventovanější protějšek, příslušný obrat přeložíme jako *vzorce v jazyce*.

Klauze *which we were unaware of before* je pro vyhledávání v IC příliš dlouhá, je třeba ji zkrátit — a ani pak nedostaneme uspokojivý počet dokladů: sousloví *we were unaware* se v IC vyskytuje třikrát: dvakrát s navazujícím *of*, jednou s *how*; čemuž odpovídají tyto tři české překlady: *nevíme o*, *jsme nevěděli*, *jsme si neuvědomovali*. Obecnější dotaz *unaware of* je v IC doložen v 401 případech, z toho v Trequ 106krát,[21] s následnými nejčastějšími protějšky:

> *o* (19), *vědom* (5), *neuvědomují* (5), *vědoma* (4), *neuvědomovala* (4), *neuvědomuje* (3), *vědomi* (3), *si neuvědomoval* (3).[22]

Obrat *My claim is (that)* je v IC doložen pouze jedinkrát, s českým protějškem *tvrdím* (*, že*); nadto nalezneme dva nominální případy bez pomocného slovesa *být* (*My claim that I'd fallen somehow* […]; […] *my claim that his programme is purely totalitarian* […]),

21 Případně se lze ptát na posloupnost lemmat *be*, *unaware*, *of*; výsledný počet výskytů je 176, na obdobný dotaz v Trequ však získáme jediný ekvivalent v jednom výskytu, totiž *vnímat*.

22 Srov. výsledky zarovnání po lemmatech: *uvědomovat* (42), *o* (12), *vědomý* (11), *uvědomovat se* (4), *vůbec uvědomovat* (3), *vnímat* (3), *vědět o* (3), *ani uvědomovat* (3), …

oba přeložené jako *moje/mé tvrzení, že*. Navzdory chabé četnosti jsou nám nabídnuta hned dvě překladatelská řešení: nominální (užitá ve všech třech strojových překladech, viz výše věty 5–7), či verbální (preferované lidskými překladateli, věty 2–4). Srov. Treqem nabízené protějšky lemmatu *claim*:

> *tvrdit* (2357, cca 25 %), *tvrzení* (1355), *nárok* (1043), *pohledávka* (738), *žádost* (600), *požadavek* (315), *žádat* (237), *prohlašovat* (232), *požadovat* (217), *prohlásit* (197), *nárokovat* (139), *námitka* (126), *vyžádat* (122), *uplatňovat* (113), *uvést* (108) …

Dotaz *this applies particularly to* našel v IC 26 dokladů (bez limitativního *particularly* 104), panuje tu opět nemalá variabilita, a to jak v užitém slovese, tak částici:

> *to se týká zejména* (7), *to se týká především* (1), *týká se zejména* (1), *zvláště se to týká* (1), *týká se to především* (1), *toto je pokud se týče o* […] *obzvlášť* (1);
>
> *to platí zejména o* (2), *to platí zejména pro* (2), *to platí zvláště o* (1), *to platí zvláště pro* (1), *to platí především pro* (1), *obzvláště to platí pro* (1);
>
> *to se vztahuje zejména na* (3), *to se vztahuje především na* (1), *se vztahují na* (1);
>
> *a zejména v* (1).

Vidíme nicméně, že kombinací dvou nejfrekventovanějších českých komponentů (*týkat se* s 12 výskyty z 26 a *zejména* s 16 výskyty z 26) dostaneme zároveň nejčastější protějšek celého sousloví, který použijeme pro své účely.

V Trequ je potřeba dotaz omezit na *applies to*, nejčastějšími nabízenými ekvivalenty jsou:

> *se* (83), *se týká* (55), *pro* (54), *platí pro* (39), *i* (14), *týká* (12), *se týká i* (9), *to* (9), *se to* (9) *týká se to* (9)

Strukturně podobné části *how languages differ* a *what they share* vyhledáváme v IC pomocí dotazů *how .* differs?*,[23] resp. *what .* shares?*. Výsledky jsou obdobné: v prvním případě dostaneme 6 dokladů (*jak se liší* 3, po 1 výskytu *že se liší*, *čím se liší*, *se lišíme*), v druhém 9 dokladů (po 1 výskytu *co sdílíme*, *co společně sdílíme*, *co sdílíš*, *co jsme všichni zdědili*, *zda se podělit o*, *se dělí o*, 1 případ je výsledkem špatného zarovnání, zbylé 2 doklady neodpovídají námi hledané struktuře). Treq na dané dotazy nepřinese žádné výsledky, hledání je třeba omezit na příslušný slovesný tvar.

Obrat *perhaps eventually* se v IC vyskytuje 6krát, pokaždé s jiným českým ekvivalentem: *nakonec snad*, *třeba nakonec*, *snad jednou*, *nakonec možná*, *až by případně*, *pravděpodobně nakonec*. Dvakrát je doložen též v Trequ: s ekvivalenty *případně* a *až případně*.

23 Pomocí regulárního výrazu ? zahrneme singulárovou i plurálovou podobu slovesa.

Z úseku *what characterises language* hledáme nevariabilní část *what characterises*, která je v IC doložena 4krát, s různými překlady: *tím, co charakterizuje*; *je specifický díky*; *co je však pro* [EU] *typické*; *charakterizuje*. V Trequ pak hledáme samostatný klíčový výraz *characterises*, jemuž nejčastěji (v 19 případech, tj. cca 40 %) odpovídá tvar *charakterizuje*; mezi 21 ekvivalenty s frekvencí 1 najdeme většinou víceslovné protějšky, nemálo z nich je použitelných: *je typický pro*, *je příznačný pro*, *je pro charakteristický*, *bývají typické pro* aj.[24]

Konečně lexém *in general* je v IC doložen cca 3 tisíci výskytů, což už je dostatečná frekvence, abychom jej s uspokojivými výsledky hledali též v Trequ.

> *obecně* (1132, cca 77 %), *všeobecně* (120), *vůbec* (51), *celkově* (33), *obecně platí* (19), *obecně vzato* (15), *všeobecně vzato* (8), *obecně řečeno* (8) …

Další inspirativní protějšky najdeme mezi ekvivalenty doloženými jednou: *obecně platí*, *celkově řečeno*, *všeobecně vzato*, *obecně řečeno*, *jako celek*, *celkově vzato* aj.

Využijeme-li výše uvedených dat k vytvoření korpusově asistovaného překladu, resp. opřeme-li se o nejfrekventovanější překladatelská řešení nabízená IC, potažmo Treqem, dostaneme tento text:

(8) *Často se říká, že prostřednictvím korpusů můžeme pozorovat vzorce v jazyce, které jsme si dříve neuvědomovali. Tvrdím, že se to týká zejména vícejazyčných korpusů. Vidíme, jak se jazyky liší, co sdílejí a nakonec snad i to, co charakterizuje jazyk obecně.*

Srovnejme jej ještě s vylepšenou, post-editovanou verzí GT (tj. nakolik tento překladač dovolil opravit prvotní podobu — větu 5):

(5a) *Často se říká, že prostřednictvím korpusů lze pozorovat vzory v jazyce, kterých jsme si dříve nebyli vědomi. Mé tvrzení je, že to platí zejména pro vícejazyčné korpusy. Můžeme vidět, jak se jazyky liší, co sdílejí a — možná nakonec — co charakterizuje jazyk obecně.*

6. ZÁVĚRY A VÝHLEDY DO BUDOUCNA

Z takto získaného překladu i procesu jeho vzniku lze vyvodit přinejmenším dva závěry:

— Překládali jsme slovo za slovo, maximálně kolokace za kolokaci, výsledek má tudíž blíže ke strojovým překladům než k těm přirozeným, lidským. Uživateli však nic nebrání daný výtvor *ad libitum* dále opracovávat, míra jeho zásahů záleží na něm. Důležitým kritériem je rovněž jeho úroveň znalostí výchozího jazyka, případně daného diskursu: zdaleka ne všechny nabízené ekvivalenty,

24 Srov. výsledky pro stejný dotaz ve verzi 1.1: *charakterizuje* (24), *typický* (3), *typické* (2), *charakteristická* (2), *charakteristický* (2), *nemluvím* (1), *charakterizují* (1), *charakterizovaný* (1), *příznačného* (1), *vyznačují* (1), *charakterizovány* (1), *staccatem* (1).

včetně těch nejčastěji doložených, jsou vhodné. Frekvenční hledisko, jež je pro uživatele Trequ primárním vodítkem, nelze absolutizovat — mnohá inspirativní řešení je možné najít i mezi ekvivalenty s frekvencí 1.

— Náš experiment měl zároveň prověřit, zda je paralelní korpus InterCorp svou aktuální velikostí (ale též složením) schopen poskytnout uspokojivé výsledky pro podobnou úlohu. Četnosti jednotlivých víceslovných jednotek převážně v řádu desítek (v Trequ často jen jednotek) by hovořily spíše proti,[25] na druhou stranu výsledný překlad považujeme za relativně kvalitní. Ovšemže záleží na míře ustálenosti daného lexému a od ní se odvíjející frekvence, ta opravdu frekventovaná slovní spojení (v našem případě diskursní částice[26] *in general*) se v Trequ přesvědčivě projeví už nyní, sousloví méně častá je třeba vyhledávat primárně v rozhraní KonText. Zásadní výhodu spatřujeme ve vzájemném propojení obou rozhraní, dovolující spojit rychlost a přímočarost (Treq) spolu s možností ověřit si přirozené chování nabízených ekvivalentů přímo v textech (KonText).

Ukázala se také užitečnost rozšíření funkcionality Trequ o víceslovné jednotky. Nutnou daní za toto rozšíření je vyšší počet zavádějících kandidátů, kteří se však na čelných pozicích frekvenčního seznamu objevují jen zřídkakdy; v případě potřeby navíc lze tuto možnost vypnout a přejít k ekvivalentům jednoslovným.

Společným jmenovatelem (a zároveň dezideratem) obou výše uvedených závěrů je přesvědčení, že další zlepšení výsledků lze očekávat úměrně s narůstajícím objemem dat,[27] větší žánrovou pestrostí textů a také s postupným zlepšováním nástrojů na automatické zarovnávání slov.

V budoucnu bychom rádi prozkoumali další možnosti zarovnání víceslovných jednotek: nabízí se např. nejprve vyhledat v textech víceslovné jednotky pomocí spe-

25 O úskalích využití paralelních korpusů, zvláště IC, v kontrastivnělingvistickém výzkumu píše výstižně M. Martínková (2014), její perspektiva se s těmi podanými námi v oddíle 2 (uživatelská, případně též lexikografická) v mnoha bodech shoduje.

26 Tuto třídu lexémů podchytává nová verze Trequ velice zdatně, viz další namátkové sondy, řazené sestupně podle počtu výskytů v Trequ:
I mean (11 374): *myslím* (4736, cca 42 %), *myslím tím* (724), *teda* (485), *tím myslím* (459), *tedy* (273), *myslím to vážně* (243), *to myslím* (238), *totiž* (213), *chci* (195), *víš* (155), *vždyť* (140), *myslím že* (134), *chci říct* (104), *no* (98), …
in fact (1242): *vlastně* (479, cca 39 %), *skutečně* (153), *popravdě* (98), *fakticky* (45), *dokonce* (36), *popravdě řečeno* (31), *naopak* (26), *opravdu* (26), *ostatně* (23), *faktem* (18), *totiž* (18), *vskutku* (15), *skutečnost* (13), …
on the other hand (75): *naopak* (24, cca 32 %), *naproti* (16), *straně* (9), *naproti tomu* (9), *zato* (8), …
what is more (7): *navíc* (3), *víc* (1), *toho* (1), *a co víc* (1), *kromě toho* (1).

27 Možným zdrojem takovýchto dat by mohl být např. paralelní korpus CzEng 1.6 (Bojar et al., 2016) o velikosti téměř 600 milionů anglických slov. Tato možnost však není samozřejmá; pro mnohé jazykové kombinace je těžké až nemožné najít větší paralelní korpusy než IC. Obecně to souvisí s větší náročností budování paralelních korpusů oproti těm jednojazyčným.

cializovaných nástrojů a následně hledat zarovnání slov už s těmito víceslovnými jednotkami; jinou eventualitou je zkusit při zarovnávání slov využít morfologickou a/nebo syntaktickou anotaci.

LITERATURA

Bojar, O. — Dušek, O. — Kocmi, T. — Libovický, J. — Novák, M. — Popel, M. — Sudarikov, R. — Variš, D. (2016): CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In: P. Sojka et al. (eds.), *Text, Speech and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12–16, 2016, Proceedings*. Berlin / Heidelberg: Springer Verlag, s. 231–238.

Cuřín, J. (1998): *Automatická extrakce překladu odborné terminologie*. Diplomová práce. Praha: Ústav formální a aplikované lingvistiky MFF UK.

Čermák, F. — Kocek, J. (eds.) (2010): *Mnohojazyčný korpus InterCorp: možnosti studia*. Praha: Nakladatelství Lidové noviny / Ústav Českého národního korpusu.

Čermák, F. — Rosen, A. (2012): The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 13, 3, s. 411–427.

Čmejrek, M. (1998): *Automatická extrakce dvojjazyčného pravděpodobnostního slovníku z paralelních textů*. Diplomová práce. Praha: Ústav formální a aplikované lingvistiky MFF UK.

Čmejrek, M. — Cuřín, J. (2001): Automatic Extraction of Terminological Lexicon from Czech-English Parallel Texts. In: *International Journal of Corpus Linguistics Special Issue 2001*, s. 1–12.

Girgzdis, V. — Kale, M. — Vaicekauskis, M. — Zarina, I. — Skadiņa, I. (2014): Tracing Mistakes and Finding Gaps in Automatic Word Alignments for Latvian-English Translation. In: A. Utka et al. (eds.), *Human Language Technologies — The Baltic Perspective. Proceedings of the Sixth International Conference Baltic HLT 2014*. Amsterdam: IOV Press BV, s. 87–94.

Chlumská, L. (2014): Není korpus jako korpus: Korpusy v kontrastivní lingvistice a translatologii. *Časopis pro moderní filologii*, 96, 2, s. 221–232.

Jirásek, K. (2011): Využití paralelního korpusu InterCorp k získávání ekvivalentů pro chorvatsko-český slovník. In: F. Čermák (ed.), *Korpusová lingvistika Praha 2011: 1 — InterCorp*. Praha: Nakladatelství Lidové noviny / Ústav Českého národního korpusu, s. 45–55.

Johansson, S. (2007): *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Kaczmarska, E. — Rosen, A. (2015): Jak najít optimální překlad polysémních jednotek — porovnání metod formální analýzy paralelních textů. *Časopis pro moderní filologii*, 97, 2, s. 157–168.

Klégr, A. — Kubánek, M. — Malá, M. — Rohrauer, L. — Šaldová, P. — Vavřín, M. (2016): Korpus InterCorp — angličtina, verze 9 z 9. 9. 2016 [online]. Praha: Ústav Českého národního korpusu FF UK. Dostupné z <http://www.korpus.cz>.

Kovář, V. — Baisa, V. — Jakubíček, M. (2016): Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography*, 29, 3, s. 339–352.

Mareček, D. (2009): Using tectogrammatical alignment in phrase-based machine translation. In: J. Šafránková — J. Pavlů (eds.), *WDS 2009 Proceedings of Contributed Papers*. Praha: Matfyzpress, s. 22–27.

Martinková, M. (2014): K metodologii využití paralelních korpusů v kontrastivní lingvistice. *Naše řeč*, 97, 4–5, s. 270–285.

Och, F. J. — Ney, H. (2003): A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29, 1, s. 19–51.

Rosen, A. (2016): InterCorp — a look behind the façade of a parallel corpus. In: E. Gruszczyńska — A. Leńko-Szymańska (eds.), *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*. Warszawa: Instytut Lingwistyki Stosowanej, s. 21–40.

Rosen, A. — Adamová, M. — Vavřín, M. (2014): Extrakce lexikálních ekvivalentů z paralelního korpusu. In: *Korpusová lingvistika Praha 2014. 20 let mapování češtiny. Abstrakty*. Praha: Ústav Českého národního korpusu, s. 177–179.

Sharoff, S. — Rapp, R. — Zweigenbaum, P. — Fung, P. (eds.) (2013): *Building and Using Comparable Corpora*. Springer-Verlag: Berlin.

Sinclair, J. M. — Mauranen, A. (2006): *Linear Unit Grammar*. Amsterdam — Philadelphia: John Benjamins Publishing Company.

Skoumalová, H. (2008): Extracting dictionaries from parallel corpora. In: F. Čermák — R. Marcinkevičienė — E. Rimkutė — J. Zabarskaitė (eds.), *Proceedings of The Third Baltic Conference on Human Language Technologies*. Kaunas: Vytauto Didžiojo Universitetas, s. 297–301.

Škrabal, M. (2016a): Paralelně — víceslovně — lépe? K možnostem nové verze databáze překladových ekvivalentů Treq. In: *Kolt 2016. Korpusy v kontrastivní lingvistice a translatologii. Abstrakty*. Praha: Ústav Českého národního korpusu. Cit. 1. 2. 2017, dostupné z <http://ucnk.ff.cuni.cz/kolt2016/KOLT2016_abstrakty.pdf>.

Škrabal, M. (2016b): *Srovnávací aspekty lotyšského a českého lexikonu: Materiály k sestavení lotyšsko-českého slovníku*. Disertační práce. Praha: Filozofická fakulta Univerzity Karlovy.

Tiedemann, J. (2000): *Automatical Lexicon Extraction from Aligned Bilingual Corpora*. Master's thesis. Magdeburg: Otto-von-Guericke-Universität.

Michal Škrabal | Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 2, 116 38 Praha 1
michal.skrabal@ff.cuni.cz

Martin Vavřín | Ústav Českého národního korpusu, Filozofická fakulta Univerzity Karlovy | nám. Jana Palacha 2, 116 38 Praha 1
martin.vavrin@ff.cuni.cz